A COMPUTERIZED MORPHO-SYNTACTIC SYSTEM OF ARABIC

Marwan Al-Bawab, Muhammad Mrayati*,
Yahya Meer Alam, and Muhammad Hassan Tayyan

Scientific Studies and Research Centre
P.O. Box 4470, Damascus
Syria

الخلاصة :

يرمي هذا النظام إلى معالجة التحليل الصرفي والنحوي للعربية باستخدام الحاسوب ، ليكون جزءاً من نظام خبير للغة العربية . وقد اقتصر في هذه المرحلة على معالجة الكلمة العربية مستقلة عن سياق النص .

يعرض هذا المقال قسمين أساسيين هما :

١ - الاشتقاق : ويجري فيه توليد جميع الكلمات العربية القياسية والسماعية انطلاقاً من جذورها الثلاثية أو الرباعية بجميع أوزانها وحالاتها الإعرابية .

٢ - التحليل : ويجري فيه تحليل الكلمة العربية صرفياً ونحوياً لتحديد سوابقها ولواحقها ، ونوعها ، وتجردها ، وزيادتها ، ووزنها ، وجذرها المشتقة منه ، وحالتها الإعرابية ، . . .

وقد أُنجز هذا العمل باستخدام الحاسوب VAX-11 780 ، وما حواه هذا المقال من نماذج هي من خرج الحاسوب نفسه .

ومما يمتاز به هذا النظام استيعابه جميع مواد العربية (الجذور) ، والقوانين الصرفية والنحوية ، إضافة إلى أن خرجه تام الضبط بالشكل (الحركات) ، وآني التنفيذ .

*To whom correspondence should be addressed.

ABSTRACT

This system is aimed at processing the morphological and syntactical analyses of Arabic using computer, so as to make it part of an Expert System for the Arabic language. At this stage the processing has been restricted to the Arabic word independently of the context.

This paper demonstrates the two main divisions of the system:

1. Generation (derivation), in which all Arabic words, standard and nonstandard, are generated right from their triliteral and quadriliteral roots, with all their balances and grammatical states.
2. Analysis, in which an Arabic word is analyzed morphologically and syntactically to determine its prefixes, suffixes, type, being bare or increased, balance, root, grammatical state, *etc.*

The work has been accomplished using the VAX-11 780 computer, and the bilingual terminal INTEGRO. The samples included (Figures and Tables) are computer output.

Peculiar to this system is its coverage of the entire gamut of the Arabic lexical entries (roots) as well as the morphological and syntactical rules. Its output is characterized by thorough administration of diacritics, and availability for instantaneous implementation.

A COMPUTERIZED MORPHO-SYNTACTIC SYSTEM OF ARABIC

INTRODUCTION

The system described in this paper is part of a set of activities undertaken by the Scientific Studies and Research Centre (SSRC) in Damascus for processing Arabic by computer. The object of these activities is to bring about an Expert System for the Arabic language, the importance of which consists in the numerous significant applications expected from it such as: language understanding, machine-aided translation, speech recognition and synthesis, teaching Arabic, and detection and correction of linguistic errors in texts [1].

The processing of Arabic using a computer is based on a number of analyses: morphological, syntactical, semantic, *etc*. [2].

The Arabic morphological system falls into two main divisions, *i.e*. word derivation (word generation) and word morphological analysis [3]. The syntactical system, on the other hand, may be processed on two levels: the first is to analyze the word independently of the context; the other is to analyze it with reference to the context.

This paper is concerned with the utilization of computers in processing the Arabic morphological system both in its analytical and generative divisions, as well as syntactical word analysis, independently of the context.

The system programs have been accomplished using a VAX-11 780 computer with the bilingual terminal INTEGRO. The SSRC is now making these programs available on IBM PC/AT personal computers, using the Turbo Prolog language.

This system is distinguished by its capacity to cover all possible entries in the Arabic language; the morphological and syntactical rules (standard and nonstandard). It is also marked by full diacritic marks, as well as instantaneous implementation of its output. The output is demonstrated in tabular forms that add to its viability for exploitation and utilization. Specimens of the computer output have been included.

ARABIC WORD DERIVATION [4]

In this section, derivatives of roots (lexical entries) are made. The increased forms of roots and the derivatives of the increased forms are also made. We have started off at the original entry (triliteral and quadriliteral roots) to find out the derivatives and increased forms.

Figure 1 represents a schema of the stages for Arabic word derivation. It establishes that the system is based on making the most of rules in the first place, so that the tables are consulted only upon the absence of rules. This in fact makes the system more adapted to becoming an expert system, and contributes towards minimizing the time for processing.

Once the letters of the lexical entry (root) have been input, the computer investigates the pertinence of the root to the lexicon. If found impertinent (*i.e*. if it is not found to be a bare verb, solid noun, or a preposition), the computer would rule it out as not a potential item of the approved lexicons. Otherwise, the computer furnishes around 40 menus enabling the utilizer (whether human or a programming or informational system) to obtain all morphological and syntactical data pertaining to the verbs and nouns derived from the input item.

It is worth pointing out here that by the term "rules" is meant the use of syntactical and morphological principles, and by the term "tables" is meant those lists of the irregular and nonstandard cases not belonging under the "rules". We should also stress the regularity of the Arabic language in most of its morphological, syntactical, and phonetic features, and also the close relation between its form and content: an attribute that places Arabic in an ideal position for processing automatically [5].

The data can be summed up as follows:

- Verb type: consonantal, vocalized, hamzated, geminated (rules) [6].
- Morphological categories of the triliteral verb: transitivity; intransitivity; in addition to transitivity letters (tables).
- Increased triliteral verbs (12 in number) and quadriliteral verbs (3) (rules + tables) [7].
- Derived nouns: standard and nonstandard, which include the active participle and its exaggeration, the passive participle, nouns of place, time and instrument, elative nouns, and qualificative adjectives (rules + tables) [8].

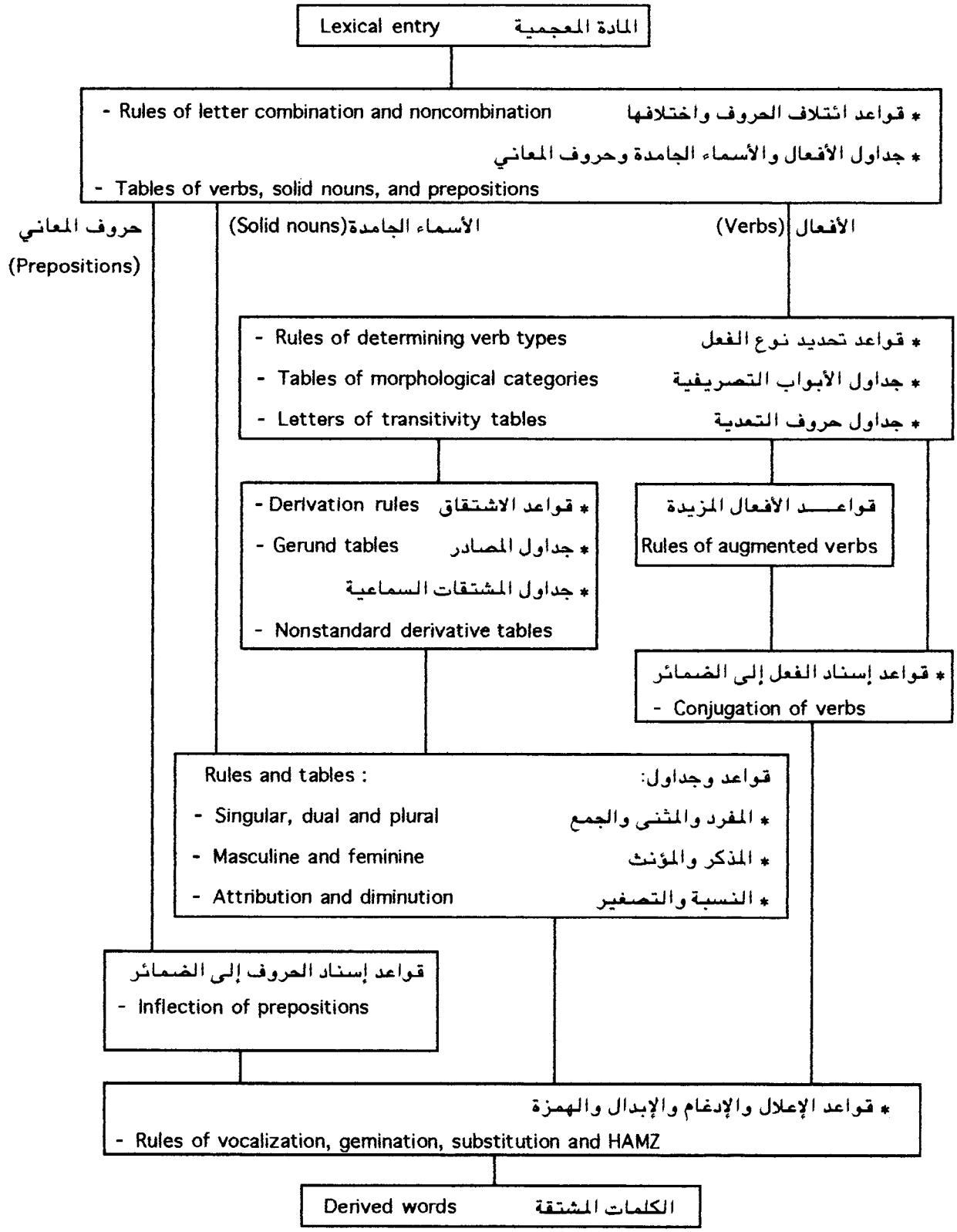
Lexical entry المادة المعجمية
- Rules of letter combination and noncombination * قواعد ائتلاف الحروف واختلافها
* جداول الأفعال والأسماء الجامدة وحروف المعاني
- Tables of verbs, solid nouns, and prepositions
حروف المعاني (Prepositions)
(Solid nouns) الأسماء الجامدة
(Verbs) الأفعال
- Rules of determining verb types * قواعد تحديد نوع الفعل
- Tables of morphological categories * جداول الأبواب التصريفية
- Letters of transitivity tables * جداول حروف التعدية
- Derivation rules * قواعد الاشتقاق
- Gerund tables * جداول المصادر
* جداول المشتقات السماعية
- Nonstandard derivative tables
قواعد الأفعال المزيدة Rules of augmented verbs
* قواعد إسناد الفعل إلى الضمائر
- Conjugation of verbs
Rules and tables : قواعد وجداول:
- Singular, dual and plural * المفرد والمثنى والجمع
- Masculine and feminine * المذكر والمؤنث
- Attribution and diminution * النسبة والتصغير
قواعد إسناد الحروف إلى الضمائر
- Inflection of prepositions
* قواعد الإعلال والإدغام والإبدال والهمزة
- Rules of vocalization, gemination, substitution and HAMZ
Derived words الكلمات المشتقة

Figure 1. Schema of Arabic-Word Derivation.

- Standard and nonstandard gerunds, nouns of one action, nouns of kind, the "mimi" and artificial nouns (rules + tables) [9].
- Conjugation of triliteral and quadriliteral verbs: active and passive, bare and increased in the past tense, in the present tense (in the "raf' ", "naṣb", "ğazm", and emphatic forms), and in the imperative mood (emphatic and non-emphatic) with all pronouns (rules) [10].
- Declension of nouns: bare and increased, solid and derived, through giving their feminine, dual, plural, attribution, diminutive forms in all cases (*i.e.* subjective, objective, and prepositional cases) (rules + tables) [11].
- Joint inflection: which covers the rules of vocalization, gemination, substitution, and hamzation. These rules apply to both verbs and nouns [12].

Samples of the Derivation System Output:

- Figure 2 illustrates the main menu of the computer on operating the derivation program. A glance at the plate demonstrates that the computer makes it possible to know readily the verb type and its morphological categories, to conjugate its active/passive voice, and to determine its derivatives, gerunds, increased forms and their derivatives. This can be accomplished by selecting the number corresponding to the desired choice.
- Figure 3 represents a secondary menu that includes the conjugation of the verb in the past tense, in the present tense "raf' ", "naṣb", "ğazm", and emphasis), and in the imperative tense (emphatic and non-emphatic). Following number selection the computer displays the relevant table of conjugation of the verb selected with all pronouns.
- Table 1 offers the conjugation of the verb (كتب) in its past tense with all pronouns. The table covers all the Arabic dependent subjective pronouns (13 in number), with the verb form linked to each. The pronouns have been arranged according to their being first person pronouns (2 pronouns), second person pronouns (5), and third person pronouns (6) respectively.
- Table 2 presents the conjugation of the verb (قال) in its imperative tense with the second person pronouns only which is the sole possibility. Note that with specific pronouns, the verb undergoes some kind of vocalization.
- Table 3 shows one standard derivative of the verb (لعب), namely the active participle. The table comprises the resultant forms of this participle with the singular (masculine and feminine), with the dual (masculine and feminine), and with the sound plural (masculine and feminine) — all in the subjective, objective and prepositional cases.
- Table 4 manifests the elative noun for the verb (كبر), associated with the definite article (ال) .

Morpho-Syntactic Analysis of the Arabic Word

In this section, the processing is performed for the words marked, wholly or partially, with diacritics and also those words not so marked. This is done by determining prefixes, suffixes, the bare and increased words, balances, roots, types, and grammatical states.

We have opted for merging the morphological and the syntactic analysis into the computer output, as this has been found of more use. Hence the joint exposition of the two analyses [13].

The computer processes the input words according to their respective parts of speech, *i.e.* verb, noun, and preposition [14]. This is because word processing independently of the context gives rise to the possibility of having a group of words overlapping nouns and verbs (*e.g.* أكرم), or prepositions and verbs (*e.g.* أنْ). However, there are a few clues that obviate such a possibility, and help to specify the part of speech to which a word appertains. For example, the nunnation mark suffixed to a word suggests that word is a noun [14]. Therefore, such words are processed neither in the verb stage nor in the preposition stage.

It should be pointed out that the analysis is at present restricted to processing Arabic and/or Arabized words.

Morpho-Syntactic Analysis of Verbs

The computer here processes the input verb, by defining its prefix, suffix, type, voice, bareness or increasedness, balance, root, grammatical state, and the pronoun linked to it.

An outline of the stages of verb analysis is shown in Figure 4. The following paragraphs give a summary of these stages.

Figure 2.

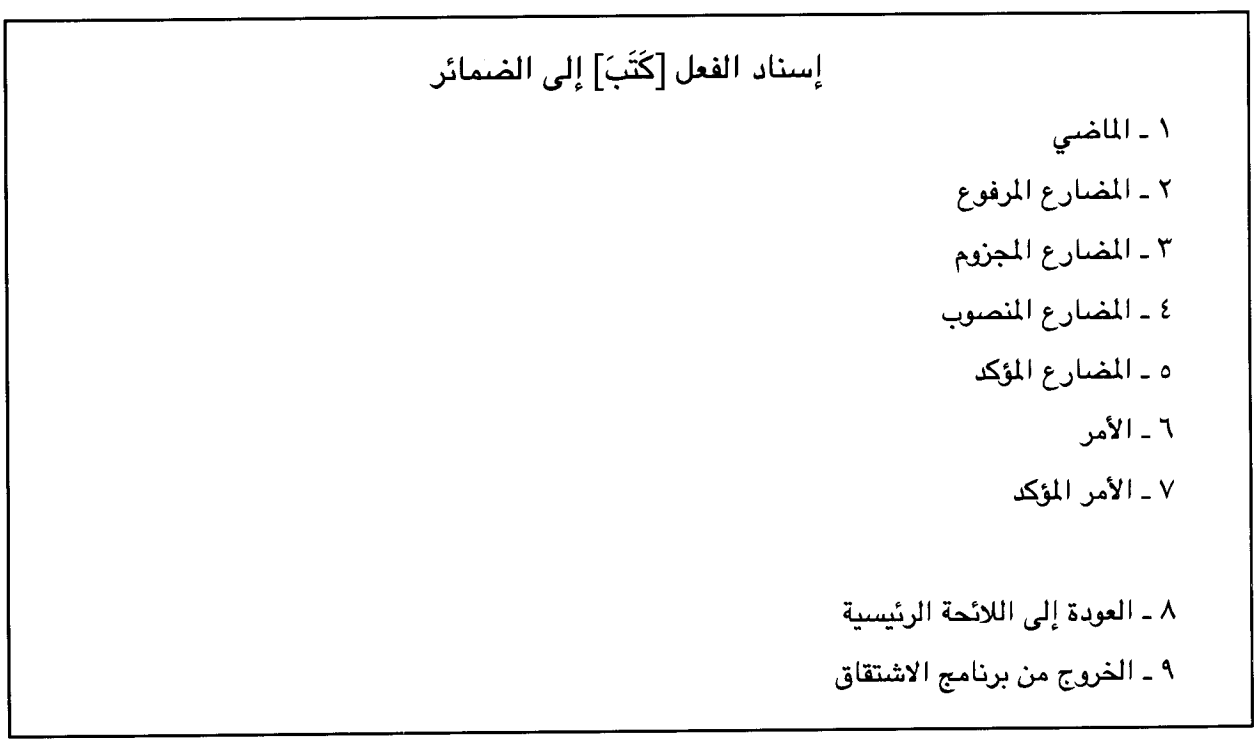

Figure 3.

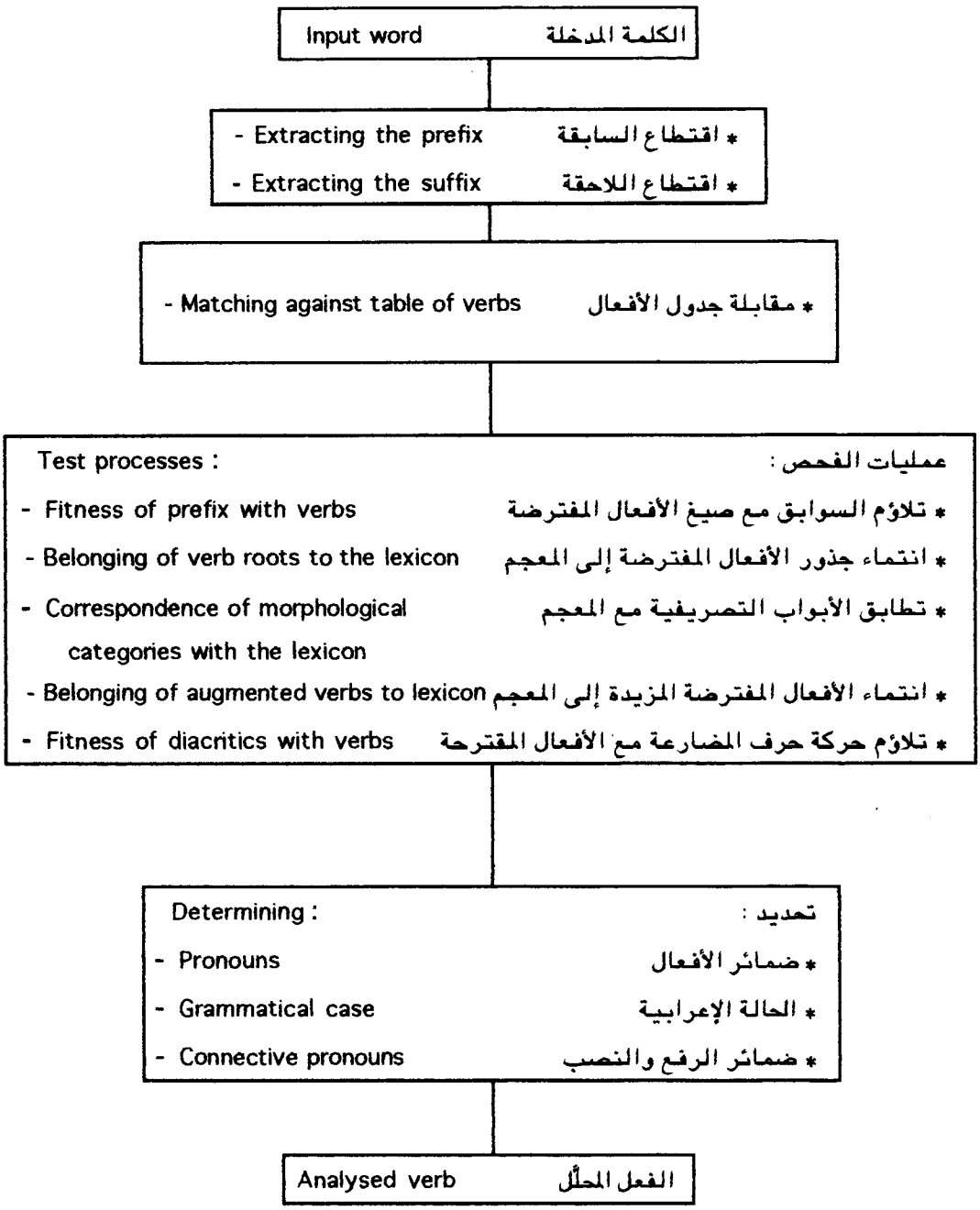
Input word
الكلمة المدخلة
- Extracting the prefix
* اقتطاع السابقة
- Extracting the suffix
* اقتطاع اللاحقة
- Matching against table of verbs
* مقابلة جدول الأفعال
Test processes :
عمليات الفحص :
- Fitness of prefix with verbs
* تلاؤم السوابق مع صيغ الأفعال المفترضة
- Belonging of verb roots to the lexicon
* انتماء جذور الأفعال المفترضة إلى المعجم
- Correspondence of morphological categories with the lexicon
* تطابق الأبواب التصريفية مع المعجم
- Belonging of augmented verbs to lexicon
* انتماء الأفعال المفترضة المزيدة إلى المعجم
- Fitness of diacritics with verbs
* تلاؤم حركة حرف المضارعة مع الأفعال المقترحة
Determining :
تحديد :
- Pronouns
* ضمائر الأفعال
- Grammatical case
* الحالة الإعرابية
- Connective pronouns
* ضمائر الرفع والنصب
Analysed verb
الفعل المحلَّل

Figure 4. Outline of Verb Analysis.

إسناد الماضي (كَتَبَ) إلى الضمائر

الضمير	تصريف الفعل	الضمير	تصريف الفعل
أنا	كَتَبْتُ	هو	كَتَبَ
نحن	كَتَبْنَا	هي	كَتَبَتْ
أنتَ	كَتَبْتَ	هما (مذ)	كَتَبَا
أنتِ	كَتَبْتِ	هما (مؤ)	كَتَبَتَا
أنتما	كَتَبْتُمَا	هم	كَتَبُوا
أنتم	كَتَبْتُمْ	هنّ	كَتَبْنَ
أنتنّ	كَتَبْتُنّ		

Table (1)

إسناد الأمر من (قَالَ) إلى الضمائر

الضمير	تصريف الفعل	الضمير	تصريف الفعل
أنا		هو	
نحن		هي	
أنتَ	قُلْ	هما (مذ)	
أنتِ	قُولِي	هما (مؤ)	
أنتما	قُولا	هم	
أنتم	قُولُوا	هنّ	
أنتنّ	قُلْنَ		

Table (2)

اسم الفاعل من (لَعِبَ)

حالة	حالة	المـذكـر	المـؤنـث
الـرفـع	مفرد	لاعِـبٌ	لاعِبَـةٌ
	مثنى	لاعِبَـانِ	لاعِبَتَانِ
	جمع	لاعِبُـونَ	لاعِبَاتٌ
النصـب	مفرد	لاعِباً	لاعِبَةً
	مثنى	لاعِبَيْنِ	لاعِبَتَيْنِ
	جمع	لاعِبِينَ	لاعِبَاتٍ
الجـر	مفرد	لاعِبٍ	لاعِبَةٍ
	مثنى	لاعِبَيْنِ	لاعِبَتَيْنِ
	جمع	لاعِبِينَ	لاعِبَاتٍ

Table (3)

اسم التفصيل المعرف بِـ (ال) من (كَبَرَ)

حالة	حالة	المـذكـر	المـؤنـث
الـرفـع	مفرد	الأَكْبَرُ	الْكُبْرَى
	مثنى	الأَكْبَرَانِ	الْكُبْرَيَانِ
	جمع	الأَكْبَرُونَ	الْكُبْرَيَاتُ
النصـب	مفرد	الأَكْبَرَ	الْكُبْرَى
	مثنى	الأَكْبَرَيْنِ	الْكُبْرَيَيْنِ
	جمع	الأَكْبَرِينَ	الْكُبْرَيَاتِ
الجـر	مفرد	الأَكْبَرِ	الْكُبْرَى
	مثنى	الأَكْبَرَيْنِ	الْكُبْرَيَيْنِ
	جمع	الأَكْبَرِينَ	الْكُبْرَيَاتِ

Table (4)

Extracting the Prefix

A verb may be prefixed by one-letter prefixes (they are 6 in number), *e.g.* the letter (ف) in (فسمع), or two-letter prefixes (there are 12 of them), *e.g.* (أو) in (أوسمع). Furthermore, each verb-tense is confined to certain prefixes. The letter (س), for instance, as a prefix goes with verbs in the present only, whereas the "hamza" goes with verbs both in the past and present, and so on.

The computer truncates the word by extracting the prefix, starting right from the longest prefix, and taking into account that these prefixes bear no relation whatever to augmented letters in increased verbs.

It merits mentioning that the extracted prefix may be inherent in the verb, *e.g.* the letter (و) in the verb (وتد). The computer, having failed to analyze (تد), replaces what has been mistaken for a prefix (*i.e.* the letter و), and processes the whole word over again.

On the other hand, the extracted prefix may be inherent in one verb and an actual prefix to another verb. For example, the letter (ل) in the word (لسعت) is inherent in the verb (لسع) and a prefix to the verb (سعى).

Extracting the Suffix

A verb may be suffixed by subjective pronouns as in: (سمعتُ), or by subjective and objective pronouns, as in: (سمعتُها), or even by one subjective and two objective pronouns, *e.g.* (سألتمونيها).

The computer truncates the word by extracting the suffix, starting from the longest suffix. It is also worth stating that, like prefixes, the extracted suffix may be inherent in the verb, *e.g.* the letter (ك) in the verb (ضحك). The computer, having failed to analyze (ضح), replaces what has been mistaken for a suffix (*i.e.* the letter ك), and processes the word all over again.

On the other hand, the extracted suffix may be inherent in one verb and actual suffix to another verb. For example, the letter (ك) in the verb (ملّك) is inherent in the verb (ملك) and a suffix to the verb (ملّ).

At this stage the compatibility between subjective and objective pronouns is also tested. The following examples are deemed incorrect for lack of compatibility: (سمعنانا), (سمعتَك).

Matching Against the Table of Verbs

In the wake of extracting the prefix and suffix, the computer matches the word with the table of verbs, so as to check it against verbs that are equal in the number of letters and diacritics. The outcome would be all the verbs presumed to match up with the given word.

The table of verbs has been arranged according to the number of letters, and in order of diacritical marks at the same time, which means that the matching process is executed in almost no time. In addition, the table of verbs includes all the data related to the verb's root, morphological category, balance, tense, and voice. This data has been allotted 7 symbols only, which indicates that the table does not require considerable computer memory.

Test Processes

The computer performs the following tests on all the presumed verbs already obtained at the previous stage. As a result the computer eliminates those verbs proved not to fulfil the requirements. The test processes aim at verifying the following:

- fitness of the prefix to the forms of the presumed verbs.
- belonging of the roots of those verbs to the computer lexicon.
- correspondence of the morphological categories of the bare presumed verbs with their counterparts in the computer lexicon.
- belonging of the increased presumed verbs to the computer lexicon.
- fitness of the diacritical sign above the letters indicating present tense (*i.e.* the letters ا, ن, ي, and ت) to the presumed-verb tense.

The afore-mentioned computer lexicon is divided into two sections: the one is for triliteral roots, the other for quadriliteral roots. Data related to triliteral roots is stored in the form of a three-dimensional matrix (28^3); each dimension indicates its respective letter of the root, so that the number in the first dimension indicates the first letter, and so on. Thus the element (12, 24, 18) in that matrix indicates the root (سمع). Say the element contains the number 4, the computer would construe the root (سمع) as containing a triliteral bare verb of the fourth category of conjugation. The same applies to all lexical data associated with every root.

In this way maximum speed in checking the relevance of a given root to the computer lexicon is secured, not to mention the fact that only limited space in the computer memory is occupied.

As for quadriliteral roots, data related to them is stored in 28 matrices, the number of each represents the first letter of the root. The remaining three letters of the root have been alphabetically arranged within those matrices, with the relevant lexical data at the side of each.

Linkage Pronouns

At this stage is determined the pronoun to which the verb is linked conformably with set rules. It is worth noting, however, that the same verb may be linked to two different pronouns. Example: the verb (تكتب) can be linked to the second person masculine singular pronoun (أنت), and to the third person feminine singular pronoun (هي).

Grammatical State of Verbs

At this stage the computer defines the verb in terms of its being "mabnī" or "mu'rab", and then defines the verb sign in either case, be it explicit or implicit, whether by diacritics, letter or omission.

An overall total of (19) such possibilities has been recorded.

At this point, too, the grammatical state of the verb is checked against the prefix. The letter (س), for instance, will never occur as a prefix to a present verb in "naṣb" (*i.e.* we do not say (سنعملَ)), neither will the letter (ل) occur as a prefix to a present verb in "raf' " (*i.e.* (ليقولُ) is incorrect).

Subjective and Objective Pronouns

These are suffixes to the verb and so they are definable by the computer, in addition to defining the letter (ت) denoting femininity and suffixed to a past verb for the third person singular, such as (سمعَتْ). Likewise, the computer defines the so-called "nūn al-wiqāya" (نون الوقاية) that must be added to a verb just before the letter (ي) for the first person singular, such as (سمعني).

Note also that the dependent pronoun (نا) may be suffixed to a verb either as a subjective pronoun (*e.g.* ضربْنا), or an objective pronoun (*e.g.* ضربَنا).

Analysis Output

The output of the morphological and syntactic analysis of the word is displayed on the computer screen or printed out in the way we are going to mention hereafter. The output may be made ready for utilization by programs or other informational systems such as the semantic system.

Examples of the Verb Analysis Output

The following are ten examples of the computer output, representing various aspects of verb morpho-syntactic analysis.

Example (1) is an analysis of the verb (خذ) which is the imperative from (أخذ), "mabnī" by the "sukūn" diacritic, and linked to the second person masculine singular (أنتَ).

Example (2) is an analysis of the present passive verb (يؤدَّب).

Example (3) is the imperative verb (اصطفِ), which is "mabnī" by omitting the vocalic letter. Notice also the substitution that has occurred.

Example (4) is an analysis of the quadriliteral increased verb (يتدحرجان) of the pattern (تفعلل).

Example (5) analyses the verb (يعِدون) which undergoes the omission of its initial radical consonant when conjugated with the present.

Example (6) is an analysis of the present verb (نرَ) in the "ğazm" by omitting the vocalic letter. It is linked to the first person plural. Notice here that the "hamza" is dropped as the verb is conjugated in the present.

Example (7) gives a two-fold analysis to the verb (وصفتْ), in which the initial (و) is once regarded as a prefix and once as an inherent letter of the verb.

Example (8) is typical for a verb that may be linked to two different persons.

Example (9) illustrates a verb with two different grammatical states.

Example (10) represents a verb derived from two different verbs.

خُذْ = فعل أمر
مجرد، من المادة (أخذ)
مبني على السكون
مسند إلى المخاطب المفرد

Example (1)

يُؤَدَّبُ = فعل مضارع، مجهول
مزيد على وزن (فَعَّلَ)، من المادة (أدب)
مرفوع بالضمة
مسند إلى الغائب المفرد

Example (2)

اصْطَفِ = فعل أمر
مزيد على وزن (افْتَعَلَ)، من المادة (صفو)
مبني على حذف حرف العلة
مسند إلى المخاطب المفرد

Example (3)

يَتَدَحْرَجَانِ = فعل مضارع، معلوم
مزيد على وزن (تَفَعْلَلَ)، من المادة (دحرج)
مرفوع بثبوت النون
مسند إلى الغائبَيْن
متصل بضمير الرفع (ا)

Example (4)

يَعِدُونَ = فعل مضارع، معلوم
مجرد، من المادة (وعد)
مرفوع بثبوت النون
مسند إلى جماعة الغائبين
متصل بضمير الرفع (و)

Example (5)

نَرَ = فعل مضارع، معلوم
مجرد، من المادة (رأي)
مجزوم بحذف حرف العلة
مسند إلى جماعة المتكلمين

Example (6)

وصَفَتْ = السابقة (وَ) + فعل ماضٍ، معلوم
مجرد، من المادة (صفو)
مبني على الفتح المقدر
مسند إلى الغائبة المفردة
متصل بتاء التأنيث الساكنة
وَصَفَتْ = فعل ماضٍ، معلوم
مجرد، من المادة (وصف)
مبني على الفتح
مسند إلى الغائبة المفردة
متصل بتاء التأنيث الساكنة

Example (7)

تَحْمِلْ = فعل مضارع معلوم
مجرد، من المادة (حمل)
مجزوم بالسكون
مسند إلى المخاطب المفرد

تَحْمِلْ = فعل مضارع، معلوم
مجرد، من المادة (حمل)
مجزوم بالسكون
مسند إلى الغائبة المفردة

Example (8)

يَغْزُونَ = فعل مضارع، معلوم
مجرد، من المادة (غزو)
مبني على السكون
مسند إلى جماعة الغائبات
متصل بضمير الرفع (ن)
يَغْزُونَ = فعل مضارع، معلوم
مجرد، من المادة (غزو)
مرفوع بثبوت النون
مسند إلى جماعة الغائبين
متصل بضمير الرفع (و)

Example (9)

نَبَتَا = فعل ماضٍ، معلوم
مجرد، من المادة (نبو)
مبني على الفتح المقدر
مسند إلى الغائبتَيْن
متصل بضمير الرفع (ا)
نَبَتَا = فعل ماضٍ، معلوم
مجرد، من المادة (نبت)
مبني على الفتح
مسند إلى الغائِبَيْن
متصل بضمير الرفع (ا)

Example (10)

Morpho-Syntactic Analysis of Nouns

The computer here processes the input noun, by defining the prefix, suffix, balance, root, grammatical case, its type in terms of derivation, whether masculine, feminine, singular, dual, or plural, and also as regards attributives and diminutive forms.

An outline of the stages of noun analysis is shown in Figure (5). A summary of these stages is given in the following few paragraphs.

Extracting the Prefix

A noun can be prefixed by one-letter prefixes (they are 7 in number), *e.g.* the letter (ب) in (بمدرسة); or two-letter prefixes (there are 15 of them), *e.g.* (ال) in (المدرسة); or three-letter prefixes (8), *e.g.* (بال) in (بالمدرسة); or four-letter prefixes (7), (وبال) in (وبالمدرسة); or five-letter prefixes (2), *e.g.* (أوَبال) in (أوَ بالمدرسة).

It is worthwhile to note that the majority of noun prefixes are different from verb prefixes.

The computer truncates the word by extracting the prefix, starting right from the longest prefix, and considering that these prefixes bear no relation to augments in increased nouns. The distinction between the "moon lām" (اللام القمرية) and the "sun lām" (اللام الشمسية), both being noun prefixes, is also observed by the computer, depending on the following letter.

It is a case in point to mention that the prefix extracted may be inherent in the noun, *e.g.* the letter (ل) in (لعِبٌ). The computer, having failed to analyze (عب), replaces what has been mistaken for a prefix, and processes the whole word over again.

Extracting the Suffix

A noun may be suffixed by prepositional pronouns such as the letter (ه) in the word (كاتبه), in addition to a number of letters connected with the sound masculine plural as in (كاتبون), sound feminine plural as in (كاتبات), dual (كاتبان), and feminine, *e.g.* (كاتبة).

The computer truncates the word by extracting the suffix, starting from the longest suffix. To be noted is the fact that, like prefixes, the suffix extracted may be inherent in the noun, *e.g.* the letter (ه) in the noun (فقيهُ). The computer, having failed to analyze (فقي), replaces what has been mistaken for a suffix, and processes the word all over again.

Fitness of Prefix to Suffix

At this stage the compatibility between the prefix and the suffix is tested. Some prefixes are never found with a noun terminating in certain suffixes. For instance definite nouns (*i.e.* with ال) which are suffixed by prepositional pronouns such as (الكتابك) are incorrect Arabic.

Matching Against the Table of Nouns

After extracting the prefix and suffix, the computer matches the word with the table of nouns, so as to check it against nouns that are equal in the number of letters and diacritics. The outcome would be all the nouns presumed to match up with the given word.

The table of nouns includes a number in the order of 450 noun balances embracing all Arabic balances.

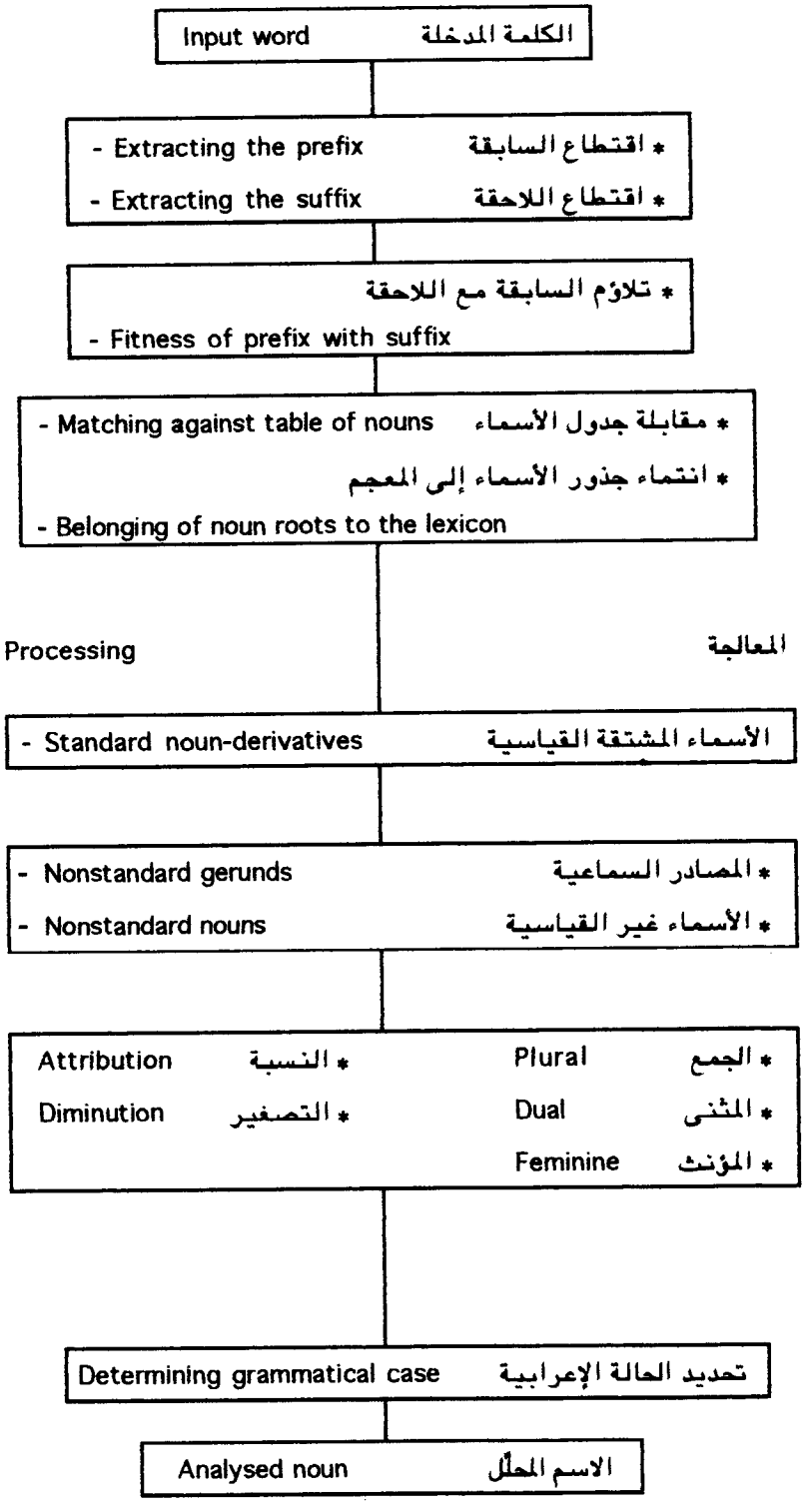
Input word الكلمة المدخلة
- Extracting the prefix * اقتطاع السابقة
- Extracting the suffix * اقتطاع اللاحقة
* تلاؤم السابقة مع اللاحقة
- Fitness of prefix with suffix
- Matching against table of nouns * مقابلة جدول الأسماء
* انتماء جذور الأسماء إلى المعجم
- Belonging of noun roots to the lexicon
Processing المعالجة
- Standard noun-derivatives الأسماء المشتقة القياسية
- Nonstandard gerunds * المصادر السماعية
- Nonstandard nouns * الأسماء غير القياسية
Attribution * النسبة
Diminution * التصغير
Plural * الجمع
Dual * المثنى
Feminine * المؤنث
Determining grammatical case تحديد الحالة الإعرابية
Analysed noun الاسم المحلّل

Figure 5. Outline of Noun Analysis.

Like that of verbs, the table of nouns has been arranged according to the number of letters, and in order of diacritical signs at the same time.

It should be emphasized here that matching a word against the table of nouns is completely independent of matching it against that of verbs, since there are words that are simultaneously nouns and verbs. The word (كاتبوه), for instance, could be looked at as the imperative verb from (كتب), or the sound masculine plural noun from (كاتبٌ). It means that, in both verb and noun analysis, the computer should presume for the input word verbs and nouns which are corresponding to it.

Belonging of Noun Roots to the Computer Lexicon

The computer at this stage seeks to ascertain that the presumed noun roots belong to the computer lexicon, and subsequently eliminates the presumed nouns of which roots are not part of the lexicon.

Processing

Noun processing has been divided into three steps:

Step 1. the computer searches for presumed nouns belonging to standard noun-derivatives, namely the active participle, the passive participle, nouns of place and time, the elative noun, in addition to certain exaggeration forms of the active participle, nouns of instrument, and qualificative adjectives.

Step 2. the computer now searches for presumed nouns pertaining to nonstandard or solid nouns, or nonstandard gerunds (*i.e.* those of bare triliteral verbs).

Step 3. in which the computer searches to find out whether the presumed nouns pertain to any of the following:

- plural forms: sound masculine or feminine, paucal or multitude plurals, plural of the plural, the form of ultimate plural, *etc.*
- dual noun forms.
- feminine forms.
- attributive forms.
- diminutive forms

Processing of presumed nouns is performed individually for each step, since the input word may be at once a standard noun-derivative and also a noun transferred from an adjective. Example: (شارب) is an active participle from (شرب), and also a noun transferred from an adjective.

The input word may also be a singular noun for one meaning and a dual noun for another. Example: (مرجان) is the dual of (مرج), and a singular masculine noun.

It could further be a singular noun in one sense, and a plural in another. Example: (حراب) is a gerund from the verb (حرب), and a multitude plural of (حربة).

Having completed the search processes, the computer defines the type of those presumed nouns which have shown sound compatibility with the input word. So the computer defines the derivative nouns, the non-derivative nouns, the plurals, duals, feminines, attributives, diminutives, *etc.* The total number of these types amounts to 49.

Grammatical Case of Nouns

Depending on its last letter(s), the word is defined by the computer in terms of "mabnī" or "mu'rab", and the mark of each, be it explicit or implicit.

An overall number of 26 grammatical cases has been recorded for the "mu'rab" nouns, of which 11 are subjective, 7 objective, and 8 prepositional cases.

Analysis Output

The output of the morphological and syntactic analysis of the word is displayed on the computer screen or printed out as follows.

Examples of the Noun Analysis Output

The following are nine examples of the computer output, representing various aspects of noun analysis.

Example (11) is an analysis of the noun (استيطان) which is an increased gerund of the pattern (استفعال), from the root (وطن).

Example (12) is an analysis of the word (جاعلون), the sound masculine plural of (جاعل).

Example (13) analyses the word (مزارع), a multitude plural.

Examples (14) and (15) offer the analysis of two nouns: an attributive and a diminutive respectively.

Example (16) illustrates a noun which is at the same time a masculine singular, and a dual of another noun.

اسْتِيطَانٌ = مصدر من الثلاثي المزيد على وزن (اسْتِفْعَال)
من المادة (وطن)
مرفوع بالضمة

Example (11)

جَاعِلُونَ = جمع مذكر سالم لِـ (جَاعِل) على وزن (فَاعِل)
من المادة (جعل)
مرفوع بالواو

Example (12)

مَزَارِعُ = جمع كثرة لِـ (مَزْرَعَة) على وزن (مَفَاعِل)
من المادة (زرع)
مرفوع بالضمة

Example (13)

تُرَابِيٌّ = اسم منسوب إلى (تُراب) على وزن (فُعَال)
من المادة (ترب)
مرفوع بالضمة

Example (14)

فُوَيْسِقٌ = اسم مصغر لِـ (فَاسِق) على وزن (فَاعِل)
من المادة (فسق)
مرفوع بالضمة

Example (15)

مَرْجَانِ = مثنى (مَرْج) على وزن (فَعْل)
من المادة (مرج)
مرفوع بالألف

مَرْجَانِ = اسم مذكر على وزن (فَعْلان)
من المادة (مرج)
مجرور بالكسرة

Example (16)

مَنَـازِلٌ = جمع كثرة لِـ (مَنْـزِل) على وزن (مَفَـاعِل)
من المادة (نزل)
مرفوع بالضمة

مَنَـازِلٌ = جمع كثرة لِـ (مَنْزِلَة) على وزن (مَفَـاعِل)
من المادة (نزل)
مرفوع بالضمة

Example (17)

جَانِحَـةٌ = مؤنث (جَـانِح) على وزن (فَـاعِل)
من المادة (جنح)
مرفوع بالضمة

جَانِحَـةٌ = اسم مؤنث على وزن (فَاعِلَـة)
من المادة (جنح)
مرفوع بالضمة

Example (18)

لاعِبُوهُ = فعل أمر
مزيد على وزن (فَاعَلَ)، من المادة (لعب)
مبني على حذف النون
مسند إلى جماعة المخاطبين
متصل بضمير الرفع (و)
متصل بضمير النصب (ه)

لاعِبُوهُ = جمع مذكر سالم لِـ (لاعِب) على وزن (فَـاعِل)
من المادة (لعب)
مرفوع بالواو وحذفت النون للإضافة
متصل بضمير الجر (ه)

Example (19)

Example (17) is the plural of two different nouns.

Example (18) represents a noun that is the feminine of a masculine noun on the one hand, and the feminine transferred from adjective, on the other hand.

Example (19) is a two-fold analysis of a word with nominal as well as verbal roots. Note in this example the dependent pronouns and their grammatical cases in both analyses of the word.

CONCLUSION

This computerized morpho-syntactic system of analysis establishes a significant move towards the construction of an Expert System for the Arabic language. The findings so far obtained are only very encouraging. We do hope the efforts of those researchers who have undertaken the task of processing Arabic by computer be promoted and properly coordinated. In fact that would be worthwhile in the interest of accelerating performance and sparing the effort and time which would otherwise be wasted, owing to perhaps unwitting and needless exploration of areas that may have already been investigated.

The SSRC will soon set out to process the syntactic system at sentence level, and thereupon to move on to processing the Arabic semantic system.

ACKNOWLEDGEMENT

The authors would like to acknowledge with gratitude the service of Mr. Said Al-Asaad of the SSRC, who has translated this paper into English.

REFERENCES

Note: For transliterating Arabic words we have adopted the International System for the Transliteration of Arabic Characters devised by the International Standardization Organization (ISO).

[1] Sāmir Attāsī, "View on the Components of Advanced Arabic Applications", *Computer Processing of the Arabic Language*, April 1985, p. 2.

[2] Yahyā Hilāl, "Al-'Ilāğ al-Āli lil-'Arabiyya wa Taṭbīqāt", *Conference Proceedings, First Kuwait Computer Conference, Kuwait, 1989*, p. 146.

[3] Reference [2], p. 145.

[4] M. Bawāb, Y. M. 'Alam, M. Ḥ. Ṭayyān, and M. Mrāyāti, "Niẓām Ištiqāq al-Kalima al-'Arabiyya bil Ḥāsib", *The Linguistics Series, Tunis, 1989*, p. 26.

[5] Nabīl 'Ali, *Al-Luğa al-'Arabiyya wa al-Ḥāsūb*. Kuwait: Ta'rīb, 1988, p. 173.

[6] Muṣṭafā al-Ğalāyīnī, *Ğāmi'ad-Durūs al-'Arabiyya*, vol. 1. Beirut: Al-Matba'a al-'Asriyya, 1962, p. 50.

[7] Reference [6], p. 218.

[8] 'Abbās Ḥasan, *An-Naḥw al-Wāfī*, 3rd edn., vol. 1. Cairo: Dār al-Ma'ārif, 1966, p. 238.

[9] Reference [8], p. 130.

[10] Muḥammad Ibn al-Qāḍī al-Qandalī, *As-Salsal: al-Madḫal fī 'Ilm Aṣ-Ṣarf*. Cairo: Dār Iḥyā al-Kutub, p. 4.

[11] Aḥmad al-Ḥamalāwī, *Šaḏa al-'Arf fī Fann aṣ-Ṣarf*, p. 66.

[12] Muḥammad Ibn al-Ḥasan al-Astrābāḍi, *Šarḥ Šāfiyat Ibn al-Hāğib*, vol. 3. Beirut: Dār al-Kutub al-'Ilmiyya, 1975, p. 66.

[13] M. Bawāb, Y. M. 'Alam, M. Ḥ. Ṭayyān, and M. Mrāyāti, "An-Niẓām aṣ-Ṣarfī an-Naḥwī lil-'Arabiyya bil-Ḥāsib", *Proceedings of the 2nd Conference on Computational Linguistics, Kuwait, 1989*, p. 24.

[14] Hafnī Nāṣif and others, *Qawā'id al-Luğa al-'Arabiyya*, p. 6.

ADDITIONAL BIBLIOGRAPHY

A. Dictionaries

Abū Ḥarb, Muḥammad Ḥayr, *Al-Mu'ğam al-Madrasī*. Damascus: Ministry of Education, 1985.

The Arab Academy, *Al-Mu'ğam al-Wasīṭ*. Cairo: Dār al-Ma'ārif, 1973.

Al-Azharī, Muḥammad ibn Aḥmad. *Tahḏīb al-Luğa*. Cairo: ad-Dār al-Miṣriyya lit-Ta'līf wat-Tarğama, 1964.

Al-Bustānī, Butrus, *Muḥīṭ al-Muḥīṭ*. Beirut: Librairie du Liban, 1983.

Al-Fārābī, Isḥāq ibn Ibrāhīm, *Dīwān al-Adab*. Cairo: The Arab Academy, 1975.

Al-Farāhīdī, Al-Ḫalīl ibn Aḥmad, *Al-'Ayn*. Qum: Dār Al-Hiğra, 1985.

Al-Fāyrūzabādī, Muḥammad Ibn Ya'qūb, *Al-Qāmūs al-Muḥīṭ*. Beirut: Mu'assasat ar-Risāla, 1986.

Ibn al-Qaṭṭā', 'Alī ibn Ğa'far, *Kitāb al-Af'āl*. Beirut: 'Ālam al-Kutub, 1983.

Al-Miṣrī, Muḥammad ibn Manẓūr, *Lisān al-'Arab*. Beirut: Dār Ṣādir.

Al-Munğid fī al-Luğa, 21st edn. Beirut: Dār al-Mašriq, 1973.

Riḍā, Aḥmad, *Mu'ğam Matn al-Luğa*. Beirut: Dār Maktabat al-Ḥayāt, 1958.

As-Saraqusṭī, Sa'īd ibn Muḥammad, *Kitāb al-Af'āl*. Cairo: The Arab Academy, 1975.

Al-Azdī, Muḥammad ibn al-Ḥasan, *Kitāb Ğamharat al-Luğa*. Beirut: Dār Ṣādir.

Az-Zubaydī, Muḥammad Murtaḍā, *Tāğ al-'Arūs*. Kuwait: At-Turāṯ al-'Arabī Series, 1965.

B. Books

Abū as-Su'ūd, 'Abbās, *Al-Fayṣal fī Alwān al-Ğumū'*. Cairo: Dār al-Ma'ārif, 1971.

Al-Astrābāḏī, Muḥammad ibn al-Ḥasan, *Šarḥ Kāfiyat ibn al-Ḥāğib*. Beirut: Dār al-Kutub al-'Ilmiyya.

Al-Ğarğānī, 'Abdul Qāhir, *Kitāb al-Miftāḥ fī aṣ-Ṣarf*. Beirut: Mu'assasat ar-Risāla, 1987.

Al-Išbīlī, Ibn 'Uṣfūr, *Al-Mumti' fī at-Taṣrīf*. 4th edn. Beirut: Dar al-'Āfāq al-Ğadīda, 1979.

An-Naḥwī, Abū Ḥayyān, *Al-Mubdi' fī at-Taṣrīf*. Dār al-'Urūba, 1982.

Qabāwa, Faḫr ad-Dīn, *Taṣrīf al-'Asmā' wal 'Af'āl*. Beirut: Maktabat al-Ma'ārif, 1988.

Paper Received 10 June 1990; Revised 4 August 1993.